Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2019) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

5th International Conference on AI in Computational Linguistics

Detecting Semantic-based Similarity Between Verses of The Quran with Doc2vec

Menwa Alshammeri[a,b,1], Eric Atwell[a], Mhd ammar Alsalka[a]

[a]*University of Leeds, Leeds, LS2 9JT, UK*
[b]*Jouf University, Sakaka, Saudi Arabia*

Abstract

Semantic similarity analysis in natural language texts is getting great attention recently. Semantic analysis of the Quran is especially challenging because it is not simply factual but encodes subtle religious meanings. Investigating similarity and relatedness between the Quranic verses is a hot topic and can promote the acquisition of the underlying knowledge. Therefore, we use an NPL method to detect the semantic-based similarity between the verses of the Quran. The idea is to exploit the distributed representation of text, to learn an informative representation of the Quran's passages. We map the Arabic Quranic verses to numerical vectors that encode the semantic properties of the text. We then measure similarity among those vectors. The performance of our model is judged through cosine similarity between our assigned semantic similarity scores and annotated textual similarity datasets. Our model scored 76% accuracy on detecting the similarity, and it can act as a basis for potential experiments and research.

Peer-review under responsibility of the scientific committee of the 5th International Conference on AI in Computational Linguistics
Keywords: The Quran, Text similarity, Semantic-based similarity, NLP, Document embeddings, Doc2vec

1. Introduction

One of the most important problems in NLP is document similarity. It has a lot of applications in many natural language processing tasks. It can be achieved using lexical similarity or semantic similarity. The semantic similarity task is computationally complex, as identifying relatedness between texts does not depend only on the conventional lexical matching methods, it goes beyond that. Such a task requires in-depth semantic analysis and domain-specific knowledge [1; 12].

[1] Corresponding author. Tel.: +44-746-067-8511.
E-mail address: scmhka@leeds.sc.uk

If two documents are semantically similar and describe the same concept, we can call them similar/ related. To determine the similarity between documents we need to define a way to mathematically measure the similarity. We also need to represent text from documents in some form of numeric representation so that we can perform similarity calculations on top of it. Hence, to measure how similar the Quranic verses are to each other semantically, we convert the documents/ verses into a mathematical object and define a similarity measure.

Our knowledge of how to represent words and sentences in a way that captures underlying meanings and relationships is rapidly expanding. Distributional semantics is computationally capable of modelling what humans do when they make similarity judgements [6; 5]. More recently, neural network-based sentence representation models have shown promising results in learning sentence embeddings [15]. In addition, AI community has provided a range of extremely powerful models that achieved state-of-the-art results in solving challenging NLP tasks. Paragraph vectors [10], or Doc2vec, is one of the most recent developments that is based on distributed representation for texts. Doc2Vec computes a feature vector for every document in the corpus. The vectors generated by doc2vec can be used for tasks like finding similarity between sentences/ paragraphs/ documents.

The Quran is a significant religious book followed by Muslims, and considered the main resource of Islamic regulations. The text encodes subtle religious meanings that are uncovered by direct and simple analysis [3]. This property of the Quran made it the right text for the purpose of analysing semantic similarity/ relatedness between individual verses, or a group of verses of the Quran [13]. The Quranic text contains details and concepts that are scattered all over its passages. The text addresses several concepts in a novel manner, as one concept extends to cover more than one sentence/verse. The same concept also may emerge in different places in the Quranic text. Consider the two verses given below:

وَقُرْءَانًا فَرَقْنَٰهُ لِتَقْرَأَهُۥ عَلَى ٱلنَّاسِ عَلَىٰ مُكْثٍ وَنَزَّلْنَٰهُ تَنزِيلًا ﴿١٠٦﴾

English Translation: And [it is] a Qur'an which We have separated [by intervals] that you might recite it to the people over a prolonged period. And We have sent it down progressively.

ٱللَّهُ نَزَّلَ أَحْسَنَ ٱلْحَدِيثِ كِتَٰبًا مُّتَشَٰبِهًا مَّثَانِىَ تَقْشَعِرُّ مِنْهُ جُلُودُ ٱلَّذِينَ يَخْشَوْنَ
رَبَّهُمْ ثُمَّ تَلِينُ جُلُودُهُمْ وَقُلُوبُهُمْ إِلَىٰ ذِكْرِ ٱللَّهِ ۚ ذَٰلِكَ هُدَى ٱللَّهِ يَهْدِى بِهِۦ مَن
يَشَآءُ ۚ وَمَن يُضْلِلِ ٱللَّهُ فَمَا لَهُۥ مِنْ هَادٍ ﴿٢٣﴾

English Translation: Allah has sent down the best statement: a consistent Book wherein is reiteration. The skins shiver there from of those who fear their Lord; then their skins and their hearts relax at the remembrance of Allah. That is the guidance of Allah by which He guides whom He wills. And one whom Allah leaves astray - for him there is no guide.

Verses[2] 17:106 and 39:23 are semantically related [13], both verses discuss how the Quran was revealed. Therefore, studying a subject must consider all related verses on that topic. Understanding, the semantic relations between the Quranic verses, can facilitate extracting meanings and concepts, and eventually presenting an insightful knowledge that helps both Muslim and non-Muslim, to understand and appreciate the Quran.

[2] We give the Arabic text with English word-by-word translation available at http://corpus.quran.com followed by Sahih International translation available at http://quran.com

Here, we examine the use a natural language processing method for detecting semantic-based similarity between the verses of the Quran. We use Doc2vec embeddings and cosine distance for similarity detection. Using Doc2vec, we scored a precision of 79% and accuracy of 76%. We scored higher than the baseline accuracy of 67%.

This paper is an effort and part of a continuing work to investigate Quranic semantic analysis and provide a solid base for potential research and development. The rest of this paper is organized as follows: Section 2 provides a survey of previous work in semantic text similarity for the Quran and Arabic text; Section3 describes experimental design for predicting semantic similarity between verses in the original Quranic text; Section 4 reports the evaluation result; Finally, Section 5 concludes and provides future directions.

2. Literature Review

In this section, we review previous work research done on the semantic text similarity on Arabic text, and in particular the Quran. The Quran has recently been regarded as a significant research subject in corpus linguistics, text analysis, and natural language processing [8; 4]. In the field of semantic similarity, many works studies semantic similarity between Arabic texts [2; 11; 14] focusing on Modern Standard Arabic, while others have concentrated on translation of the Holy Quran. One significant work conducted on the Arabic text of the Holy Quran is Qursim [13]. The work presents a broad dataset where semantically similar or related verses are linked together. Qursim is a large corpus of 7,600 pairs of related verses from the Quran. They used lexical similarity-based approaches like Term Frequency- Inverse Document Frequency (TF-IDF) to improve their results. Their experiments showed only 869 of the 7,679 pairs shared common words.

Another research published in [1] used a lexical similarity-based technique to compute text similarities in the Arabic text of the Holy Quran. Using the TF-IDF technique and normalization, it aimed to produce verses from the Holy Quran that are identical or relevant to a user's given query verse. It also used an N-gram and a machine learning algorithm to classify chapters (Surahs) as Makki or Madni (LibSVM classifier in Weka[3]). Only common main words are used to compute similarity in this study. The lack of semantic-based similarity search is a major flaw in this study. Efforts have also been made on the Holy Quran's translated version. [9] established a Question Answering System based on a single chapter from the Holy Quran, the Cow Chapter (Surah Baqarah). The authors classified the output to minimize the number of insignificant results returned. Fasting and Pilgrimage verses from Surah Baqarah were classified using neural networks.

[12] conducted a thorough investigation into the similarities between sacred texts. This research was conducted on sacred Bible and Holy Quran texts. To extract features and compute similarity between the documents, various statistical methods were used. A variety of distance scales were used, including Euclidean, Hillinger, Manhattan, and Cosine. The study looked at overall similarities between two documents based on their topics. It does not go any further in terms of comparing sentences from different texts. In this work we utilize natural language processing algorithm to capture semantic-based similarity between all the verses of the Quran in the original classical Arabic.

3. Experimental Design

The objective of this experiment is to use Doc2vec method to predict if pairs of verses are related; share the same meaning. To measure how similar the Quranic verses are to each other semantically, we convert the documents/ verses into a mathematical object and define a similarity measure. We build a Doc2vec model and train it on the original

3 https://www.cs.waikato.ac.nz/ml/weka/

Quran corpus. We test our model for predicating similarity using test dataset created from Qursim corpus. The experiment is composed of multiple stages: preparing the data, model training and generating embeddings, computing verses similarity and results. Each of these stages is discussed in the following sub-sections.

3.1. The Data

For the purpose of training and testing our model, we created annotated datasets using existing scholarly resource. The new dataset is a CSV file that contains 9315 pairs of related and nonrelated verses. The dataset contains 3079 pairs of verses that are related; imported from Qursim [4] dataset. We picked pairs that are related with a strong relationship; degree of relevance of 2. The dataset also contains 6236 pairs of verses that are nonrelated; randomly generated to be not in Qursim and have a degree of relevance of 0. The file contains nine columns; five of them are imported from the original Qursim dataset. In Qursim, each pair of verses <ss:sv, ts:tv> are related with a degree of relevance 0, 1, or 2. The other four columns are created for the sake of the experiment, and they are Verse1, Verse2, vid1, and vid2. The Verses text are imported from the Arabic Original Quran dataset (Tanzil documents)[5]. We use the Verse text without Diacritics to facilitate the training process. The dataset columns are described in table 1 along with examples of the data in table 2.

Table 1: Description of the Dataset columns

Column #	Column name	Description
1	ss	Surah(chapter) Id of the first item in the pair
2	sv	Verse Id of the first item in the pair within chapter ss
3	ts	Surah (chapter) Id of the second item in the pair
4	tv	Verse Id of the second item in the pair within chapter ts
5	relevance	Degree of relevance which could be 0 or 1
6	Verse1	Verse 1 text
7	Verse2	Verse 2 text
8	Vid1	Id for each verse in column 7
9	Vid2	Id for each verse in column 8

Table 2: Samples of the test dataset

vid1	ss	sv	Verse1	vid2	ts	tv	Verse2	relevance
183	2	2	ذلك الكتاب لا ريب فيه هدى للمتقين	184	10	57	يا أيها الناس قد جاءتكم موعظة من ربكم وشفاء لما في الصدور وهدى ورحمة للمؤمنين	2
309	2	11	وإذا قيل لهم لا تفسدوا في الأرض قالوا إنما نحن مصلحون	310	8	73	والذين كفروا بعضهم أولياء بعض إلا تفعلوه تكن فتنة في الأرض وفساد كبير	2

3.2. Generating Vectors using Doc2vec

Now, we need to transform our text documents (verses of from the dataset) into a numerical, vectorized form, which can later be used to calculate the cosine distance between two different verses to determine how semantically

[4] The Qursim corpus contains 7683 pairs of related verses that are created from the original Quranic text, and collected from scholarly sources, totalling up to 15,366 documents (verses).

[5] https://tanzil.net/docs/resources

similar they are, or by the clustering algorithm to group similar documents together. We use doc2vec to generate the verses embeddings.

Le and Mikolov in [10] have proposed paragraph vectors, or Doc2vec; an unsupervised method for learning distributed representation for pieces of texts. They show that their method captures many documents semantics in dense vectors and can be used for different downstream tasks [7]. Doc2vec generates vector representations of variable-length pieces of text, such as sentences, paragraphs, or documents, using a neural network approach. These vector representations have the advantage of capturing the meanings of the input texts and their context. This means that texts with similar meanings or contexts would be closer in vector space than texts with different meanings or contexts. We used gensim to train a Doc2vec model on our corpus and create vector representations of the Quranic verses. We build the model and train it on the Arabic text of the Quran using the original Arabic text from Tanzil project.

4. Model Training

In order to train a doc2vec model, the training documents need to be in the form "TaggedDocument", which basically means each document/verse receives a unique id. Only the documents that are used for training purposes should be tagged. Before feeding the verses to the model, we need to pre-process them. We separated each verse into different words (tokenization) and formed list of words for each of them along with the tagging. Because of ambiguity created when applying stemming, the presentation of documents was affected negatively by stemming. The count of words in each verse should not be affected by stop words; see table 3 for details. Therefore, we decided against removing any stop words and stemming. We trained the model and fine-tuned hyper-parameters. We experimented with different models using different settings of hyperparameters to find the optimal values for these parameters.

Table 3: Processing the input with and without stop words

Removing stop words, 99% of the verses include **51** or fewer words.
Keeping stop words, 99% of the verses include **51** or fewer words.

5. Comparing individual documents using Cosine similarity

To inspect relationships between documents numerically, we calculate the cosine distances between their inferred vectors. Cosine 6Distance/Similarity - It is the cosine of the angle between two vectors, which gives us the angular distance between the vectors; one of the most common and effective ways of calculating similarities. Therefore, we developed a function that takes as its parameters the doc2vec model we just trained and the two documents to be compared. As a measure of the documents' similarity, the function then returns a value between 0 and 1, where the larger the value, the more similar the documents. We iterated through each of the verses pair in the dataset and found out what is the cosine Similarity for each pair.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise.cosine_similarity.html

6. Results and Evaluation

Initially, we inspected our model by inferring new vectors for unseen documents/ verses from the Quran and used them to evaluate our model. For each verse we got the most similar verse from the trained model along with the similarity score, which indicates the model was properly trained. Here are few examples.

Example 1:

Train Document (410):

«يا أيها الذين آمنوا لا تتخذوا بطانة من دونكم لا يألونكم خبالا ودوا ما عنتم قد بدت البغضاء من أفواههم وما تخفي صدورهم أكبر قد بينا لكم الآيات إن كنتم تعقلون»

O you who have believed, do not take as intimates those other than yourselves, for they will not spare you [any] ruin. They wish you would have hardship. Hatred has already appeared from their mouths, and what their breasts conceal is greater. We have certainly made clear to you the signs, if you will use reason.

Similar Document (3436, 0.8588156700134277):

«ضرب لكم مثلا من أنفسكم هل لكم من ما ملكت أيمانكم من شركاء في ما رزقناكم فأنتم فيه سواء تخافونهم كخيفتكم أنفسكم كذلك نفصل الآيات لقوم يعقلون»

He presents to you an example from yourselves. Do you have among those whom your right hands possess any partners in what We have provided for you so that you are equal therein [and] would fear them as your fear of one another [within a partnership]? Thus do We detail the verses for a people who use reason.

Example2:

Train Document (3621):

«فأعرضوا فأرسلنا عليهم سيل العرم وبدلناهم بجنتيهم جنتين ذواتي أكل خمط وأثل وشيء من سدر قليل»

But they turned away [refusing], so We sent upon them the flood of the dam, and We replaced their two [fields of] gardens with gardens of bitter fruit, tamarisks and something of sparse lote trees.

Similar Document (1037, 0.9169043898582458):

«وأمطرنا عليهم مطرا فانظر كيف كان عاقبة المجرمين»

And We rained upon them a rain [of stones]. Then see how was the end of the criminals.

6.1. Predicting similarity

We tested the model capability to predict if two verses are related or not based on their cosine similarity. Our dataset contains 9315 pairs of verses that are either labelled with 1 if related, and 0 if non-related. We computed the cosine similarity for each pair in the dataset by applying the cosine similarity on the associated vectors. Using a threshold of 0.60 for the cosine similarity, we consider the pairs with similarity equal or above the threshold to be related (1), otherwise non-related (0). The value (0-1) is the cosine similarity score to determine if a pair of verses are similar/ related or not. We then compared the actual results (1 or 0 per the annotation in the dataset) with the predicated ones (1 or 0 per the similarity score).

To evaluate our model's performance, we use the Accuracy[7] metric; proportion of prediction the model classified correctly. Accuracy can be good to establish some sort of a baseline. In this case, 67%[8] will be our baseline for

[7] Accuracy = (#True positives + # True negatives) / total number of pairs

[8] We compute the baseline accuracy as: number of actual nonrelated pairs (TN) / total number of pairs; TP =0

accuracy. Using Doc2vec, we scored higher than the baseline. Our model scored 76% accuracy, 79% precision, and F1-score of 51%. We calculated a confusion matrix report and related statistics and printed the results.

Table 4: Confusion matrix and classification report on the model performance

	Precision	Recall	F1 Score	Support
0	0.75	0.95	0.84	6236
1	0.79	0.37	0.51	3079
Accuracy			**0.76**	9315

	Actual pairs	
Predicted	TP	FP
	5937	299
	TN	FN
	1151	1928

Label	Actual	Predicated
0	6236	7865
1	3079	1450

The graphs below show the distribution of verses based on their similarity. It is evident that Doc2vec computes around 47% of the pairs of verses to be similar (1450), which is around half the real distribution of the similar pairs (3097).

Figure 1: The distribution of verses pairs based on cosine similarity

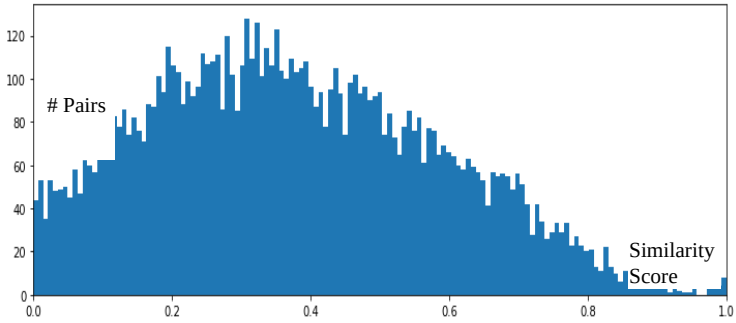
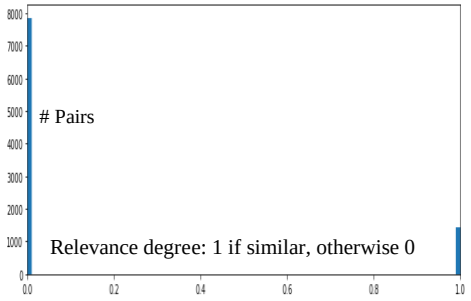


Figure 2: The distribution of verses pairs based on predicted similarity (0: non-related, 1: related)



7. C**onclusion and Future work**

We presented a natural language processing model that can be used to predict semantic-based similarity between the verses of the Quran in the original Arabic text. Using Doc2vec, we were able to locate semantically related verses. We obtained 76% accuracy, 51% F1-score. Our results can stand as a baseline for potential experiments.

The paper is an endeavor within an ongoing work towards learning the semantic relations between the Quran's verses. The ultimate goal is building a comprehensive knowledge-base for the analysis of the Quranic verses and motivating potential research and development. In the future, we plan to experiment with different similarity measures,

other than cosine similarity, that could potentially obtain deeper semantic-based similarity. We also plan to exploit the power of emerging deep learning models such as ELMo and BERT and compare their performance.

Acknowledgements

The first author is supported by a PhD scholarship from the Ministry of Higher Education, Saudi Arabia. The author is grateful for the support from Jouf University for sponsoring her research.

References

[1] Akour, M., Alsmadi, I. M., & Alazzam, I. (2014). MQVC: Measuring quranic verses similarity and sura classification using N-gram. WSEAS Transactions on Computers, vol. 13, pp. 485-491, 2014.

[2] Alian, M., & Awajan, A. (2018, November). Arabic semantic similarity approaches-review. In 2018 International Arab Conference on Information Technology (ACIT) (pp. 1-6). IEEE

[3] Alqahtani, M., & Atwell, E. (2016, June). Arabic Quranic search tool based on ontology. In International Conference on Applications of Natural Language to Information Systems (pp. 478-485). Springer, Cham.

[4] Atwell, E. S., Dickins, J., & Brierley, C. (2013). Natural Language Processing Working Together with Arabic and Islamic Studies. Engineering and Physical Sciences Research Council (EPSRC). EP/K015206/1. Online. Accessed: 29.06.2014. http://gow.epsrc.ac.uk/NGBOViewGrant.aspx?GrantRef=EP/K015206/1

[5] Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. The journal of machine learning research, 3, 1137-1155.

[6] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3, 993-1022.

[7] Dai, A. M., Olah, C., & Le, Q. V. (2015). Document embedding with paragraph vectors. arXiv preprint arXiv:1507.07998.

[8] Dukes, K. The Quranic Arabic Corpus. 2009-2011 [cited 2021 15 March]; Available from: https://corpus.quran.com/

[9] Hamed, S. K., & Ab Aziz, M. J. (2016). A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification. J. Comput. Sci., 12(3), 169-177.

[10] Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In International conference on machine learning (pp. 1188-1196). PMLR.

[11] Mahmoud, A., Zrigui, A., & Zrigui, M. (2017, April). A text semantic similarity approach for Arabic paraphrase detection. In International conference on computational linguistics and intelligent text processing (pp. 338-349). Springer, Cham.

[12] Qahl, S. H. M. (2014). An Automatic Similarity Detection Engine Between Sacred Texts Using Text Mining and Similarity Measures. Thesis. Rochester Institute of Technology. Accessed from https://scholarworks.rit.edu/theses/8496/

[13] Sharaf, A. B. M., & Atwell, E. (2012, May). QurSim: A corpus for evaluation of relatedness in short texts. In LREC (pp. 2295-2302).

[14] Schwab, D. (2017, April). Semantic similarity of arabic sentences with word embeddings. In Third arabic natural language processing workshop (pp. 18-24).

[15] Wang, S., Zhang, J., & Zong, C. (2016). Learning sentence representation with guidance of human attention. arXiv preprint arXiv:1609.09189.